Hadoop Framework to Identify Crop Diseases and Yield Prediction

Prashant Dilip Kamthe[1] Prof. Nagaraju Bogiri[2]
[1,2]Department of Computer Engineering
[1,2]KJ College of Engineering and Management Research, India

Abstract— In the growth of Information Technology, Big data come forth as a blazing topic. The main source of human survival depends on agriculture; where it needs a key contribution in the field of crop data analysis. This paper gives a purpose about how to find experiences from accuracy agriculture information through big data approach. In this way, gathering the valuable data in an effective way drives a framework towards major computational challenges in crop analysis where information is remotely gathered. For the storage purpose of huge data availability in agriculture, we are intending Hadoop framework for our work to store a huge volume of crop data. This work gives a better prediction for the farmers to plant which kind of crops to their farm field based on their soil content to improve the productivity. The random forest algorithm is integrated with the MapReduce programming model in Hadoop framework. Data is collected, clean and normalized. The Data is collected from laboratory reports, websites etc. then cleansing of data is done that is main information is extracted from unstructured redundant data. This Normalized data is uploaded on Hadoop Distributed File system and save it in supported form of Hive. HiveQL used to analyze the agricultural data. It is a SQL like query language and by analyzing crop disease symptoms, it finds out disease name and provide a solution based on evidence from historical data. This result is in form of graphs that will helpful for recommending a solution that is high symptoms similarity.

Keywords: Big Data Analytics, Hadoop, Hive, HiveQl Hobbits, MapReduce

I. INTRODUCTION

In precision agriculture, real time and historically generated data is collected in structured and unstructured datasets. The world population increasing day by day and it is expected to reach 10 billion in next 35 years. To feed the world population, an essential development in the agricultural production and disaster management are now important. It is essential to gather data from various sources of agriculture for predictive decisions. The gain in agricultural production and management are depending on weather conditions. So the prediction of area based weather is an essential task for future farming in agriculture.

Apache Hadoop and Hive are often used for finding the issues of agriculture with facilitate of huge data analytics. Apache Hadoop is a platform comes below the distributed environment. System is developed that determine diseases and suggest solutions based on historical information. And this framework can facilitate researchers in deciding and it's easily comprehensible .The solution that's extremely used for a selected illness has highest priority. For demonstration developed framework is used to spot Paddy crop leaf blast illness and suggest a solution. In computer science, the definition of data is information in raw (such as alphabets, numbers, or symbols), which might be processed by computer sure enough significance. Today, digitization is extremely common within the world. It converses analog data to digital type, which might be way more simply keep, processed and transmitted. Digital data and knowledge give an expensive source of enormous scale accumulated data. Is digital information simply accustomed share and read? Will digital data be converted into helpful knowledge with awareness value? Lots of analysis establishments realize that digital information database can become an excellent treasure, and after all, society wants are by no means satisfy with the restricted data sharing. Thus data mining becomes a new valuable research direction. a large quantity of data are created altogether walks of life daily, and the unit of data measuring develops from byte, KB, MB, GB, TB to PB, EB, ZB, YB. Extracting valuable information and exploitation data for decision-making are the core functions that we tend to want to achieve, solely during this way can data produce value. Data acquisition is not any longer a technical drawback within the age of big data, but how will we tend to explore the inherent law when face such vast and sophisticated data.

II. REVIEW OF LITERATURE

In this paper [1], a Big Data environment for agricultural soil analysis from computed tomography (CT) images is proposed. Their work done through three layers: source, Big Data environment, and applications. The application involves the statistical analysis, reconstruction, and visualization of the 3D image and soil analysis. The samples obtained from the source are processed in the Hadoop framework. The framework provides the better understanding about the soil and the prediction leads to detect the problems occurring in the soil.

In this paper [2], a Qualitative data analysis using regression method for agricultural data builds an information system for both the customer and the farmer. This utilizes the regression analysis to project the crop production which fetches the data from the cloud storage to limit the cost and easy to access. The data mining techniques such as Classification, Association rule mining, and the regression are discussed in this. These provide a qualitative analysis of data for the agricultural field.

In this paper [3], classification of large datasets using the random forest algorithm in various applications. The applications such as network intrusion detection, email spam detection, gene classification, credit card fraud detection, text classification. The framework of random forest classifier is to rank the importance of variables to avoid the classification problems, proximity, out of bag error with the features.

In this paper [4], a decision analysis for the dynamic crop rotation model with Markov process concept is proposed. The Markov process leads to taking final decision about the crop which includes the factors such as price and production. This provides the probability of the producers to choose to continue farming even there is a small production. The Markov process still has a disadvantage on handling the large dataset easily leads to getting more traffic.

All rights reserved by *www.ijsrd.com*

In this paper [5], rice crop yield prediction in India using support vector machines is the proposed technique. The parameters considered for this study is precipitation, temperature, evapotranspiration, area, production, and yield. The support vector machine approach (SVM approach) is a supervised machine learning technique which is mainly used for processing the set of labeled training set data. The performance is analyzed using the Sequential minimal optimization (SMO) classification. However, this can handle only the labeled data for the classification or regression purpose.

III. SYSTEM OVERVIEW

The proposed work is intended to help the farmers and re-searchers to understand the crop field and to choose a suitable crop for it. Various parameters are considered from soil to atmosphere for predicting the suitable crop. Soil parameters such as type, ph level, iron, copper, manganese, sulphur, organic carbon, potassium, phosphate, nitrogen. Climate such as temperature, precipitation, and solar energy and topography is considered on hill station with the parameter slope and elevation. By considering the entire parameter in agricultural field benefits the prediction process in a better way. The random forest algorithm is used to classify the dataset which provides result in good accuracy with poor error rate. Since this framework can handle large dataset by processing it in MapReduce programming model.

The phase of the proposed work

1) Data Collection
2) Data Classification(Random Forest Algorithm)
3) Mathematical Model
4) Hadoop Framework MapReduce programming model
5) Final Prediction

A. Data Collection

The dataset can be gathered in two ways: One is fetching data from the online resource and the other is from the sensor node data. The online datasets are readily available in the form of data matrix or single database table. The dataset may

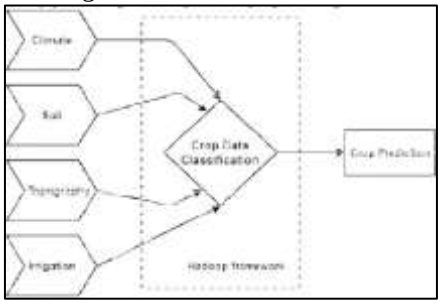

Fig. 1: The Dataset parameters to be considered for predicting the suitable crops

Contain more data which are in structured or unstructured formats. Since some dataset is highly secured need some initial registration and authentication process to access. The other form of data is sensor datasets, which mainly used by an organization/environment. The organization/environment has a sensor node infrastructure where the data are collected as per their programming framework and store it in the backend (Cloud Storage).

B. Random Forest Algorithm

The collected data are classified based on random forest algorithm. The random forest algorithm is a classification algorithm based on ensemble classifier. The training sets (t) are generated from the original dataset which is further used to build the decision tree. Each and individual training sets is constructed as the decision tree, finally, a random forest is generated. Each sample of the testing dataset is predicted by the entire decision tree.

$$t = (ai, bj) where i = 1, 2, ...n and j = 1, 2, ...m \quad (1)$$

Sampling the (t) training datasets is the initial process carried out in the original dataset. Then the decision tree is constructed for each training sets in the random forest model. The steps carried out in random forest algorithm is (a)

1) Consider the training dataset t as mentioned earlier, and the number of training dataset be n, and the number of classifiers be M.
2) The input variables t(m) number is utilized to determine the decision at a node of the tree.
3) Using the bootstrap sample, the training set is selected with the replacement from all N. The remaining cases are utilized to estimate the error of the tree by predicting their classes.
4) For each node, randomly choose the base node from t h e decision. Estimate the best split m variables in the training set.
5) Each tree is fully grown.

C. Mathematical model

Let S be the Hadoop Framework. S=N,L,D,O,P,F,E
Identify the inputs as
N = i k i is total no of Data Node
L = i i is the link between two Data sets D=i ki is Datasets
Identify the output as O O=
Identify the processes as P = F1, F2,F3,F4 F1( ) :: Fetching data from Sensor Node. F2( ) :: Convert data into Key, Value. F3( ) :: Reduce and Analysis data. F4( ) :: Identify Diseases. F5( ) :: Generate Predication. Identify failure case as F =
F = p k p no Data has successfully Created from node Identify success case (terminating case) as E Success is defined as
E = p k p highest-priority node receives the data packet successfully
S=N,L,D,O,P,F,E

D. Hadoop Framework MapReduce programming model

A Classical approach in Hadoop framework consists of loading data into the HDFS and the MapReduce programming model is used for processing the dataset. The MapReduce programming model works on both the structured and the unstructured data. The map task is done by mapper class and the reduce task is done the reducer class. The mapper class generally tokenizes the input data to sort it and this will act as the input for the reducer task. Reducer class removes the duplicate entries from the key-value pair. The primary characteristics of the MapReduce is to merge the input ¡key, value¿ pair with the list of ¡key, value¿ pair. The input is processed by mapper function to produce the intermediate set of ¡key, value¿ pair. The reducer merger the output produced from mapper to form a smaller set of values.

All rights reserved by www.ijsrd.com

E. Final Prediction

The final stage is the suitable crop based on the prediction obtained from the dataset containing the parameter such as climate, soil, topography, irrigation. The word data is plural, not singular.

IV. RESULTS AND DISCUSSION

The confusion matrix illustrates the accuracy of the solution to a classification problem. The classification result can be obtained from the true positive, false positive, true negative and false negative rate. The term true and false refer to whether the predicted corresponds to the trusted external judgments. Consider P as positive instances and N as negative instances then the true positive rate and false positive rate is TPR and FPR respectively.

$$TPR=TP/P=TP/(TP+FN)$$

$$FPR=FP/N=FP/(FP+TN)$$

Precision is defined as the number of true positives (TP) over the number of true positives plus the number of false positives (FP).

$$P=TP/(TP+FP)$$

Recall is defined as the number of the true positives (TP) over the number of true positives plus the number of false negatives (FN).

$$R=TP/(TP+FN)$$

The F1 score is a measure of the tests accuracy where both the precision and recall are used to compute the rate.

$$F1 = 2x[(PXR)/(P+R)]$$

Accuracy is defined as the number of all correct predictions- divided by the total number of the dataset. In exact, 500 data are collected from the sensor node. The dataset was divided into two sets as training sets and testing test as 325 and 175 respectively.

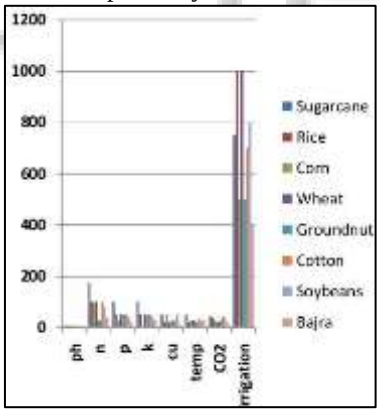

Fig. 2: The prediction of crop from the parameter such as ph, n, p, k, cu, temp, Co2, irrigation

	Predicted Positive	Predicted Negative
Observed Positive	121	08
Observed Negative	07	39

Table I: Confusion Matrix with the Positive and the Negative Value for both True and False.

Recall =0.9453 Precision=0.9380 Accuracy =0.9143 F1 score=0.9416

V. CONCLUSION

Agriculture is the backbone of many countries including India. Since integrating the information technology with the agriculture will guide the farmer to improve the productivity. In this proposed work a framework described works faster and gives better accuracy in prediction than the current system to predict the suitable crop for the field. It includes various parameters from soil and atmosphere to analyze the crop by processing it in Hadoop framework. This prediction makes the farmers to improve the productivity, growth, and quality of the plants. The future research will be devoted to predicting the required fertilizer and pesticide ratio based on the atmospheric and soil parameters for the farm land.

ACKNOWLEDGMENT

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We are thankful to the authorities of Savitribai Phule University, Pune and concern members of cPGCON2018 conference, organized by, for their constant guidelines and support. We are also thankful to the reviewer for their valuable suggestions.

REFERENCES

[1] Gabriel M. Alves, Paulo E. Cruvinel, Big Data environment for agri- cultural soil analysis from CT digital images, published in Semantic Computing (ICSC), IEEE Tenth International Conference,Feb 2016.
[2] allavi V. Jirapure, Prof. Prarthana A. Deshkar Qualitative data analysis using Regression method for Agricultural data, published in Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), World Conference, 29 Feb/March 2016.
[3] Mohammed Zakariah, Classification of large datasets using Random Forest Algorithm in various applications: Survey published in International Journal of Engineering and Innovative Technology (IJEIT), Volume 4, Issue 3, September 2014.
[4] Tyrone T. Lin, Chung-Shiao. Hsieh, A Decision Analysis for the Dy- namic Crop Rotation Model with Markov Processs Concept published in Industrial Engineering and Engineering Management (IEEM), IEEE International Conference, 10-13 Dec. 2013.
[5] Niketa Gandhi, Leisa J. Armstrong, Owaiz Petkar, Amiya Kumar Tripa- thy, Rice Crop Yield Prediction in India using Support Vector Machines, published in Computer Science and Software Engineering (JCSSE), 13th International Joint Conference, 13-15 July 2016.
[6] Snehal S.Dahikar1, Dr.Sandeep V.Rode, Agricultural Crop Yield Predic- tion Using Artificial Neural Network Approach published in International Journal Of Innovative Research In Electrical, Electronics, Instrumentation And Control Engineering, Vol. 2, Issue 1, January 2014.
[7] Wu Fan, Chen Chong, Guo Xiaoling, Yu Hua, Wang Juyun, Prediction of crop yield using big data, Published in: Computational Intelligence and Design (ISCID), 2015 8th International Symposium, 12-13 Dec.2015.

All rights reserved by *www.ijsrd.com*

[8] Snehal S. Dahikar, Sandeep V. Rode, Pramod Deshmukh, An Artificial Neural Network Approach for Agricultural Crop Yield Prediction Based on Various Parameters, published in International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Volume 4, Issue 1, January 2015.
[9] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, The Hadoop Dis- tributed File System, published on Mass Storage Systems and Technolo- gies (MSST), IEEE 26th Symposium, 3-7 May 2010.

All rights reserved by *www.ijsrd.com*